Ein binaurales Richtungshörsystem für mobile Roboter in echoarmer Umgebung

A Binaural Sound Localization System for Mobile Robots in Low-reflecting Environments

Jürgen Adamy, Kyriakos Voutsas und Volker Willert

In diesem Beitrag wird ein binaurales Lokalisationsmodell für einen autonomen mobilen Roboter vorgestellt, das sich am menschlichen Hörsystem orientiert. Im Mittelpunkt steht die technische Umsetzung der neuronalen Verarbeitungsmechanismen der Olivenkerne des auditorischen Systems, die ein binaurales Richtungshören ermöglichen. Das System ist in der Lage, im Bereich des Sichtfeldes des Roboters eine Schallquelle in echoarmer, d. h. reflexionsarmer Umgebung anhand von Laufzeitunterschieden und Pegeldifferenzen zu lokalisieren. Die Möglichkeit des Richtungshörens verbessert die Wahrnehmung der unmittelbaren Umgebung des Roboters entscheidend und erlaubt z. B. die Ausrichtung des Roboterkopfes zur Schallquelle.

This paper presents a biologically inspired binaural sound localization model for autonomous mobile robots. It mainly focusses on the technical realisation of the neural networks of the superior olivary complex, which extract binaural cues. The system is able to localize one acoustic source in the field of view of the robot in a low-reflecting environment on the basis of interaural time and level differences. The possibility of sound localization improves the spatial cognition of the robot within the surrounding area to be more humanlike.

Schlagwörter: Binaurales Hörsystem, biologisch basiertes System, mobiler Roboter, neuronales Netz

Keywords: Binaural sound localization, biologically inspired system, mobile robot, neural network

1 Einleitung

Die Wahrnehmung unserer Umwelt und die zwischenmenschliche Kommunikation hängen stark von einem funktionierenden Hörvermögen ab. Das menschliche Hörsystem hält eine Menge Rekorde im menschlichen Körper. So ist es in der Lage, den eintreffenden Schall so fein zu differenzieren, dass wir allein durch Hören unterscheiden können, ob heißer oder kalter Kaffee in eine Tasse gegossen wird [1]. Außerdem ist es das schnellste Sinnessystem des Menschen und dient unter anderem als primäres Alarm- und Lokalisationssystem. Der Mensch kann mit ihm eine Schallquelle dreidimensional im Raum und sogar einzelne Schallquellen in einem Geräuschteppich orten.

Will man wissen, wie der Mensch das Problem der Lokalisation von Signalquellen löst, muss man verstehen, wie die Signalverarbeitung im menschlichen auditorischen System funktioniert. Ausgehend von verschiedenen neurophysiologischen Experimenten ist heute bekannt, welche Wahrnehmungsunterschiede der Mensch zur Bestimmung des Ortes von Signalquellen im Raum benutzt. Dabei spielt neben der monauralen Lokalisation vor allem die binaurale Lokalisation eine entscheidende Rolle. Über die neuronalen Mechanismen, die diese Wahrnehmungsunterschiede extrahieren, lassen sich bis heute größtenteils nur Vermutungen anstellen. Es sind hauptsächlich die Funktionen, nicht die Funktionsweisen der für die Lokalisation verantwortlichen neuronalen Verarbeitungszentren bekannt [2].

Die Lokalisation erfolgt anhand von Zeit- (ITD: Interaural Time Difference) und Intensitätsunterschieden (ILD: Interaural Level Difference) des akustischen Signals zwischen beiden Ohren, wie in Bild 1 verdeutlicht. Dabei spielt

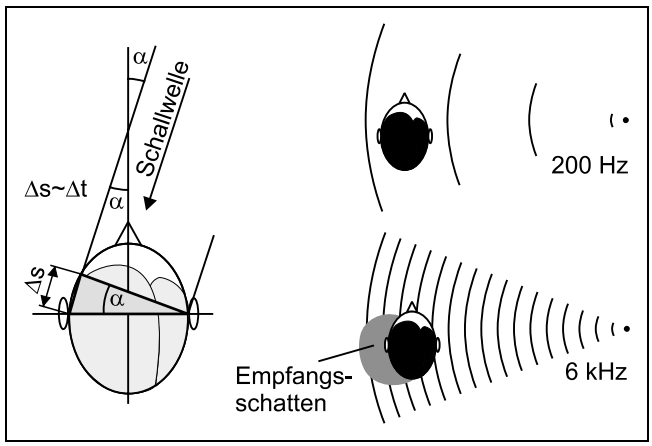

Bild 1: Interaurale Zeit- und Pegeldifferenzen.

auch die Signalfrequenz eine Rolle: kurzwellige Signale besitzen bei asymmetrischem Einfall einen Empfangsschatten auf der abgewandten Kopfseite. Sie führen daher zu höheren Pegeldifferenzen zwischen den Ohren als langwellige Signale.

Das Außenohr bewirkt dabei eine richtungsabhängige Filterung des Schallsignals. Die Filterwirkung beruht auf der Modifikation der Schallwellenausbreitung durch Dämpfung, Beugung, Reflexion und Resonanz der Schallwellen. Dabei spielen die geometrischen bzw. anatomischen Formen des Kopfes und der Schulter, sowie der Einfluss der Ohrmuschel eine entscheidende Rolle.

Binaurales menschliches Hören kann mit einer Anordnung von zwei Mikrofonen als ,,Ohrersatz“ nur bedingt nachgebildet werden. Erst durch die Berücksichtigung der akustischen Filtereigenschaften des Kopfes und der Ohren sowie der neuronalen Gehirnstrukturen werden Aufnahmen möglich, die dem natürlichen Hören am nächsten kommen. Im Gehör und Gehirn erfolgt eine äußerst komplexe Signalverarbeitung, die die Amplitudenverteilung, die spektrale Zusammensetzung und die zeitliche Struktur des akustischen Signals erfasst.

Es gibt bereits einige binaurale Simulationssysteme [3–6], welche mit Hilfe von Head-Related Transfer Functions (HRTFs) und Laufzeit- sowie Pegelunterschieden arbeiten. Die Güte binauraler Richtungshörsysteme ist jedoch noch verbesserungswürdig.

Im Folgenden wird ein bionisch inspiriertes binaurales Modell vorgestellt, das die Fähigkeit des Richtungshörens des menschlichen auditorischen Systems zum Teil nachbildet. Dabei wird zum einen auf den aus [7; 8] bekannten Stereausis Algorithmus für die Auswertung von Laufzeitdifferenzen Δt zwischen beiden Ohren zurückgegriffen. Zum anderen wird für die Auswertung von Pegeldifferenzen der Comparausis Algorithmus eingeführt. Die Ergebnisse beider Algorithmen werden durch ein Probabilistisches Neuronales Netz (PNN) kombiniert, das die Richtung der Schallquelle bestimmt. Ausgelegt ist das System für das Richtungshören von Roboterköpfen; einer bisher selten behandelten Thematik.

2 Biologisches Vorbild

Unsere Ohrmuschel bündelt den (eintreffenden) Schall. Von dort wird er über den Gehörgang auf das Trommelfell und mit Hilfe der Gehörknöchelchen, dem Hammer und dem Amboss, zum Innenohr weitergeleitet (Bild 2). Ausgelöst durch die Schwingungen des Steigbügels entstehen Wanderwellen in der Cochlea. Diese breiten sich über die in der Cochlea befindliche Basilarmembran aus, auf der die Rezeptorzellen (Haarzellen) sitzen. Die Basilarmembran besitzt ortsabhängige mechanische Eigenschaften und bremst daher die Wanderwellen vergleichbar mit Brandungswellen ab. Sie brechen nach Aufsteilen aufgrund von Reibungskräften rasch zusammen. Hohe Frequenzen werden sehr schnell, niedrige entsprechend langsamer abgebremst. Diese Ortsabhängigkeit der Frequenz befähigt das Ohr zur Frequenzanalyse. Während hohe Frequenzen nur die Haarzellen an der Basis erreichen, laufen tiefe Frequenzen weit in die Cochlea. Demnach arbeitet das Innenohr wie eine Filterbank.

Nachdem das Schallereignis den Weg des Schalltransportes durch Außen- und Mittelohr sowie die Transformation in elektrische Nervenimpulse im Innenohr durchlaufen hat, schließt sich, wie in [2; 9; 10] erklärt, der kompliziertere Teil des Hörvorgangs an, die neuronale Verarbeitung. Die topographische, auch tonotop genannte Abbildung der Frequenzen in der Cochlea wird über den Hörnerv auf die nächste Hörbahnstation, den Nucleus Cochlearis (NC), und sukzessiv auf alle weiteren Hörbahnstationen bis zum Hörkortex räumlich geordnet übertragen (Bild 3).

Die für das Richtungshören wichtigsten Stationen davon sind die oberen Olivenkerne, MSO, LSO und MNTB. Das – in verschiedene Frequenzkanäle zerlegte und durch Intervallabstände von Aktionspotentialen zeitlich kodierte – neuronal vorliegende Signal erreicht den NC über den Hörnerv. Die morphologisch und physiologisch unterscheidbaren Neuronen des NC haben ganz unterschiedliche Kodierungseigenschaften. Die so genannten Buschzellen (AVCN) werden über wenige sehr große Synapsen von jeweils wenigen Nervenfasern innerviert. Sie geben die Information der Nervenfasern nahezu unverändert weiter an die

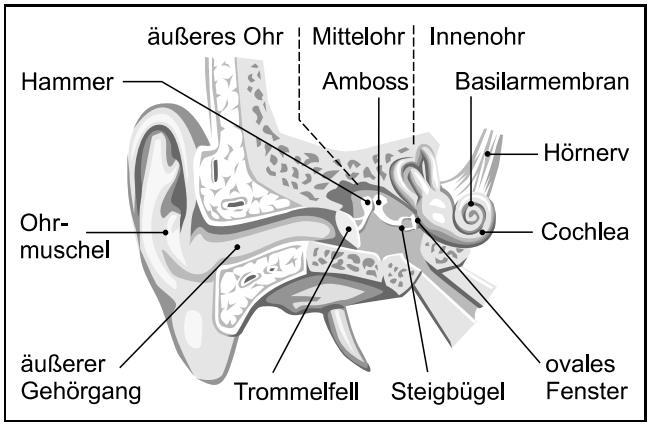

Bild 2: Das Ohr.

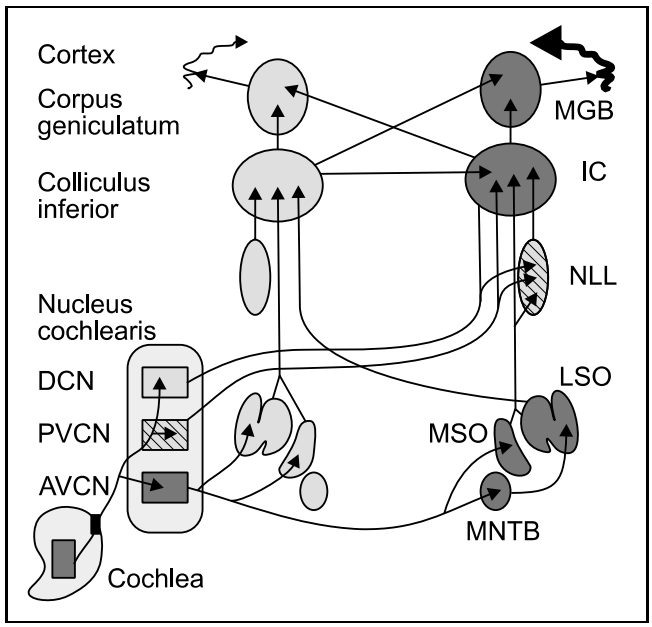

Bild 3: Schaltschema des Gehörsinnes (nach [10]).

Neuronen der Olivenkerne. Diese können die Stärke und die Zeitpunkte der Aktivierung von Neuronen des linken und rechten NC vergleichen. Die Funktionen der anderen Neuronentypen in den Zentren PVCN und DCN im NC sind für die weitere neuronale Verarbeitung in den oberen Olivenkernen nicht von Bedeutung.

Die oberen Olivenkerne transformieren die monauralen Informationen beider Ohren zu binauralen Informationen über Zeit- und Intensitätsunterschiede. Diese werden an das Hörsystem des Mittelhirns weitergegeben. Die ersten beiden folgenden Neuronengebiete sind zum einen der Inferior Colliculus (IC), und zum anderen der Nucleus Lateralis Lemniscus (NLL). Es ist bis heute unklar, welche Rolle die beiden neuronalen Verarbeitungsstationen bei der weiteren Transformation des Ausgangs der oberen Olivenkerne spielen [11]. Es gibt jedoch erste Untersuchungen über die räumliche neuronale Kodierung von Frequenz und Modulation im IC [12]. Die Existenz einer tonotopischen Kodierung in den Olivenkernen ist zwar bewiesen, der Zusammenhang zwischen Kodierung und interauralen Zeit- und Intensitätsunterschieden ist aber nicht geklärt. Bis jetzt ist experimentell belegt, dass die verschiedenen Bereiche der oberen Olivenkerne – mittlere obere Olive (MSO), seitliche obere Olive (LSO) und mittlerer Kern des Trapezkörpers (MNTB) – auf unterschiedliche binaurale Wahrnehmungsunterschiede reagieren (Bild 3). Weiterhin wird spekuliert, dass die weitergegebene Information von MSO und LSO im IC separat abgelegt wird [2; 11].

Fazit ist, die Funktion – das „Was“ – von MSO und LSO ist bekannt, die Funktionsweise – das „Wie“ – ist unklar. Abgesehen davon sind einige Modelle für die mögliche neuronale Kodierung in der MSO entwickelt worden. Das älteste und am weitesten verbreitete Modell ist das Coincidence-Model von Jefress [13]. Ein weiteres biologisch inspiriertes Modell ist der Stereausis Algorithmus von Shamma [7], der im Kapitel 4 benutzt wird. Zuerst wird im Kapitel 3 jedoch ein Modell der Schallverarbeitung im Innenohr beschrieben.

3 Modelle der Schallverarbeitung im Innenohr

Das biologische Vorbild des auditorischen Systems des Menschen soll den technischen Anforderungen entsprechend in ein diskretes mathematisches Modell umgesetzt werden. Daher sind die Schallsignale mit einer Abtastperiode T, hier $T = 1/44\,100\,\mathrm{Hz}$, zu diskretisieren. Die Implementierung beschränkt sich dabei auf die biologischen Stationen der Signalverarbeitung, welche für die Lokalisation von Schallquellen verantwortlich sind. Dabei interessiert bei der Verarbeitung der Schallwellen bis zum Hörnerv im Wesentlichen nur die Umsetzung des Innenohres. Dieser Teil der menschlichen Signalverarbeitung kann nahe an der Realität modelliert werden.

Das verwendete Cochleamodell von Patterson [14] basiert auf einer Reihe von Bandpassfiltern, den so genannten ERB-Filtern. Die „Equivalenten Rechteckigen Bandbreiten“ (ERB) sind eine psychoakustische Bezeichnung für die Bandbreite jedes Filters der auditorischen Filterbank, gemessen an jedem charakteristischen Ort entlang der Cochlea. Ein ERB-Filter modelliert das entstehende Signal in einer einzelnen Nervenzelle des Hörnervs bzw. einem einzelnen Kanal der Filterbank (Bild 4).

Jeder einzelne Bandpassfilter ist ein diskreter Filter 8. Ordnung, der aus vier Gammatone Filtern 2. Ordnung zusammengesetzt ist. Die Implementierung erfolgt nach [15] und hat die folgende Form in der z-Ebene:

$$
\begin{aligned}
F_i(z) = & \frac{-Tz^2 + Te^{-b_i T}\left(\cos\left(\omega_{c_i} T\right) + \sqrt{3+2^{\frac{3}{2}}}\sin\left(\omega_{c_i} T\right)\right) z}{-z^2 + 2e^{-b_i T}\cos\left(\omega_{c_i} T\right) z - e^{-2b_i T}} \\
& \cdot \frac{-Tz^2 + Te^{-b_i T}\left(\cos\left(\omega_{c_i} T\right) - \sqrt{3+2^{\frac{3}{2}}}\sin\left(\omega_{c_i} T\right)\right) z}{-z^2 + 2e^{-b_i T}\cos\left(\omega_{c_i} T\right) z - e^{-2b_i T}} \\
& \cdot \frac{-Tz^2 + Te^{-b_i T}\left(\cos\left(\omega_{c_i} T\right) + \sqrt{3-2^{\frac{3}{2}}}\sin\left(\omega_{c_i} T\right)\right) z}{-z^2 + 2e^{-b_i T}\cos\left(\omega_{c_i} T\right) z - e^{-2b_i T}} \\
& \cdot \frac{-Tz^2 + Te^{-b_i T}\left(\cos\left(\omega_{c_i} T\right) - \sqrt{3-2^{\frac{3}{2}}}\sin\left(\omega_{c_i} T\right)\right) z}{-z^2 + 2e^{-b_i T}\cos\left(\omega_{c_i} T\right) z - e^{-2b_i T}} .
\end{aligned}
$$

Dabei ist i die Nummer des Filterkanals, ω_{c_i} die charakteristische Frequenz und b_i die feste Bandbreite einer bestimmten charakteristischen Frequenz.

Den Verlauf des aufgesplitteten Signales in den einzelnen Kanälen der Filterbank, aufgetragen über der Zeit, kann man mit einem so genannten Cochleagramm graphisch veranschaulichen. In Bild 5 ist ein solcher Verlauf für ein Signal, bestehend aus zwei überlagerten Sinusschwingungen (300 Hz und 3000 Hz), und einer Filterbank mit

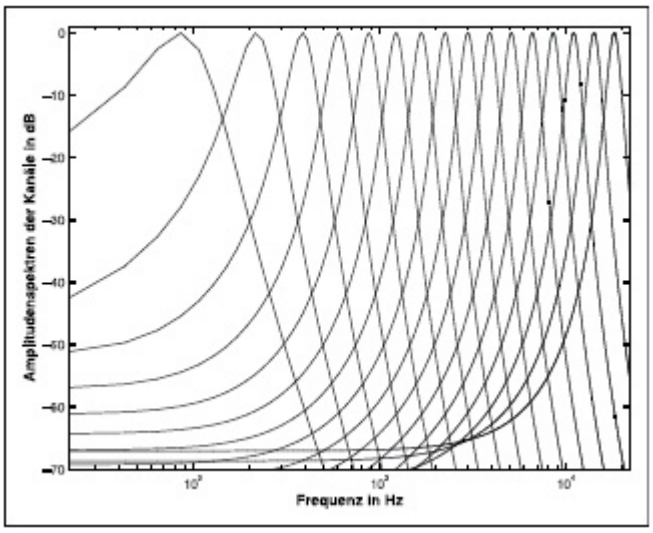

Bild 4: Amplitudengänge einer 16-Kanal Filterbank.

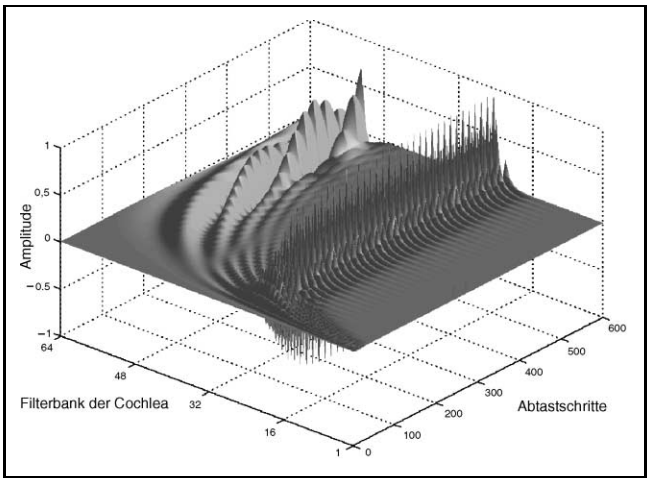

Bild 5: Cochleagramm (linkes Ohr) zweier Sinusschwingungen (300 Hz, 3000 Hz), die zum Abtastschritt $k = 0$ mit 0 beginnen.

$n = 64$ Kanälen bzw. Filtern in einem Frequenzbereich von 50–13 500 Hz dargestellt. Hierbei entspricht die höchste Kanalnummer (64) dem Kanal mit der niedrigsten charakteristischen Frequenz (50 Hz). Dies ist deshalb so gewählt worden, weil tiefe Frequenzen weit in die Cochlea vordringen.

Wegen der begrenzten Geschwindigkeit der Wanderwellen in der Cochlea, können bei benachbarten Kanälen zeitverzögerte Versionen der Antwort eines bestimmten Frequenzkanals aufgenommen werden. Die zwei Kanäle der Filterbank mit den charakteristischen Frequenzen 300 Hz und 3000 Hz in Bild 5 filtern die zwei Sinusschwingungen. Die benachbarten Kanäle, deren Bandbreite mit diesen zwei Kanälen überlappt, filtern die Sinusschwingungen leicht zeitverzögert. Damit entstehen in der Cochlea zeitverzögerte Versionen des gleichen Signals. Diese Eigenschaft der Cochlea wird im folgenden Kapitel für die Implementierung des Lokalisationsmodells benutzt.

4 Modelle der neuronalen Verarbeitung

Die hier vorgestellten Modelle für die einzelnen neuronalen Stationen sind bionisch inspirierte Umsetzungen, die stellvertretend für die bis jetzt bekannten Funktionen der neuronalen Verarbeitungszentren modelliert werden. Wie im Kapitel 2 angesprochen, sind die Olivenkerne für das Richtungshören des menschlichen Hörsystems verantwortlich, wobei die „mittlere obere Olive“ (MSO) Informationen über die Laufzeitunterschiede und die „seitliche obere Olive“ (LSO) Informationen über die Pegelunterschiede zwischen den Ohren kodieren. Der Stereausis Algorithmus von Shamma [7; 8] ist ein biologisches Funktionsmodell dieser neuronalen Struktur. Er wird hier zur Kodierung interauraler Laufzeitunterschiede verwendet. Diese Entwurfsidee wird mit dem hier eingeführten Comparausis Algorithmus aufgegriffen, um auch interaurale Pegelunterschiede tonotopisch kodiert darzustellen.

Das Stereausis Netzwerk ist ein mögliches neuronales Modell für die MSO. Dieses Netzwerk benötigt keine Verzögerungslinien wie das Jefress-Modell [13] um Laufzeitunterschiede festzustellen. Es benutzt die Verzögerungen, die in den Wanderwellen der Basilarmembran, d. h. den Gammatone Filtern, vorhanden sind (Kapitel 3), um die Kreuzkorrelationsfunktion zwischen den beiden Innenohrausgängen näherungsweise zu berechnen. Damit detektiert das zweidimensionale Modell die momentanen Unterschiede der Auslenkungen der Basilarmembrane der Cochleae beider Ohren.

Die Kanalausgänge der als Filterbank modellierten Cochlea werden in das Netzwerk (Bild 6) geleitet, das daraus ein 2D-Bild erzeugt. Die Achsen des Bildes repräsentieren die charakteristischen Frequenzen der Filterbänke von linkem und rechtem Innenohr. Die Elemente des Bildes, c_{ij}, werden berechnet, indem der momentane Ausgang x_i des i-ten Frequenzkanals des rechten Ohres mit dem momentanen

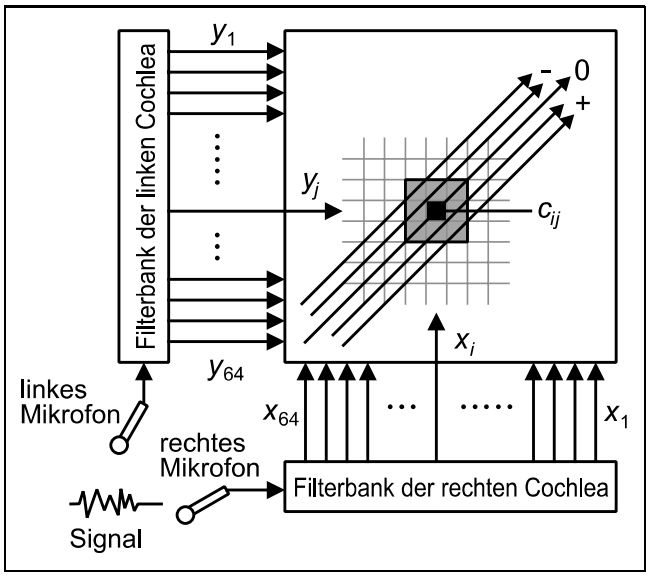

Bild 6: Stereausis Netzwerk zur räumlichen Darstellung der Kreuzkorrelation zwischen den Ausgängen der Filterbänke beider Innenohre.

Ausgang y_j des j-ten Frequenzkanals des linken Ohres multipliziert wird. Multipliziert man über eine gewisse Anzahl m von Zeitpunkten ($m = 1764$ in dieser Arbeit) und mittelt die Ergebnisse, dann entsteht aus dem Stereausis Netzwerk das so genannte Stereagramm mit den Elementen:

$$c_{ij} = \sum_{k=1}^{m} x_i(k) y_j(k), \quad i, j = 1, \ldots, n\,. \tag{1}$$

Damit entsteht eine $n \times n$ Matrix, wobei n die Anzahl der Filter jeder Filterbank ist. Für $j \in M = [i-3, i-1] \cap [i+1, i+3]$ wird näherungsweise die Kreuzkorrelation der Signale x_i und y_j für $\tau = \pm 1, \pm 2, \pm 3$ berechnet:

$$c_{ij} = \sum_{k=1}^{m} x_i(k) y_j(k) \approx \hat{s}_{xy}(\tau) = \sum_{k=1}^{m} x_i(k) y_i(k+\tau)\,. \tag{2}$$

Diese Näherung gilt, denn für $j \in M$ liefern die Gammatone Filter, wie gesagt, zeitverzögerte Antworten der Filter mit $j = i$. Die Elemente c_{ij} der Hauptdiagonalen des Stereagramms beinhalten dagegen alle exakten Kreuzkorrelationsergebnisse von $x_i(k)$ und $y_j(k)$ mit $\tau = 0$. Für $j \notin M$ beinhalten die Kanäle x_i und y_j keine zeitverzögerten Versionen des gleichen Signals. Deswegen gibt Gl. (2) die Kreuzkorrelation dieser Signale für $\tau > 3$ oder $\tau < -3$ nicht mehr wieder.

Bild 7 zeigt zwei Stereagramme für unterschiedliche Einfallswinkel α.

Wenn man das Stereausis Netzwerk mit den Ausgängen von nur einer Cochlea speist, so stellt das Stereagramm die räumliche Autokorrelation der Schallwelle dar, welche an dem entsprechenden Ohr ankommt. Es ist also möglich, die Autokorrelationen c_{ij}^x und c_{ij}^y beider Ausgänge der Cochleafilterbänke räumlich darzustellen. Die zeitliche Autokorrelation an der Stelle $\tau = 0$ ist proportional zur Leistung und damit zum Pegel des Signals. Vergleicht man die Autokorrelationen beider Cochleafilterbänke, so bekommt man ein Maß für die Pegelunterschiede zwischen beiden, an linkem und rechtem Ohr eintreffenden Schallwellen. Einen praktikablen Vergleich der Pegelunterschiede aller Kanäle liefert das logarithmische Verhältnis d_{ij} beider Pegel der übereinstimmenden Kanäle von linker und rechter Cochlea:

$$d_{ij} = 10 \cdot \log \frac{c_{ij}^x}{c_{ij}^y} = 10 \cdot \log \frac{\hat{s}_{xx}(0)}{\hat{s}_{yy}(0)}, \quad \text{mit} \quad i = j\,. \tag{3}$$

Ähnlich wie beim Stereagramm entsprechen die Elemente d_{ij} für $j \in M = [i-3, i-1] \cap [i+1, i+3]$ näherungsweise der Autokorrelation. Das entstehende Bild wird im weiteren als Comparagramm und die Abfolge der Rechenschritte als Comparausis Algorithmus bezeichnet. Das entsprechende Netzwerk ist in Bild 8 zu sehen.

Im Grunde genommen ist das Comparagramm nichts anderes als das logarithmische Verhältnis der Elemente der Koinzidenzmatrizen beider Filterbankausgänge. Vergleicht man die zwei in Bild 9 und 10 dargestellten Comparagramme so ist, je nach Einfallswinkel des Schalles, ein

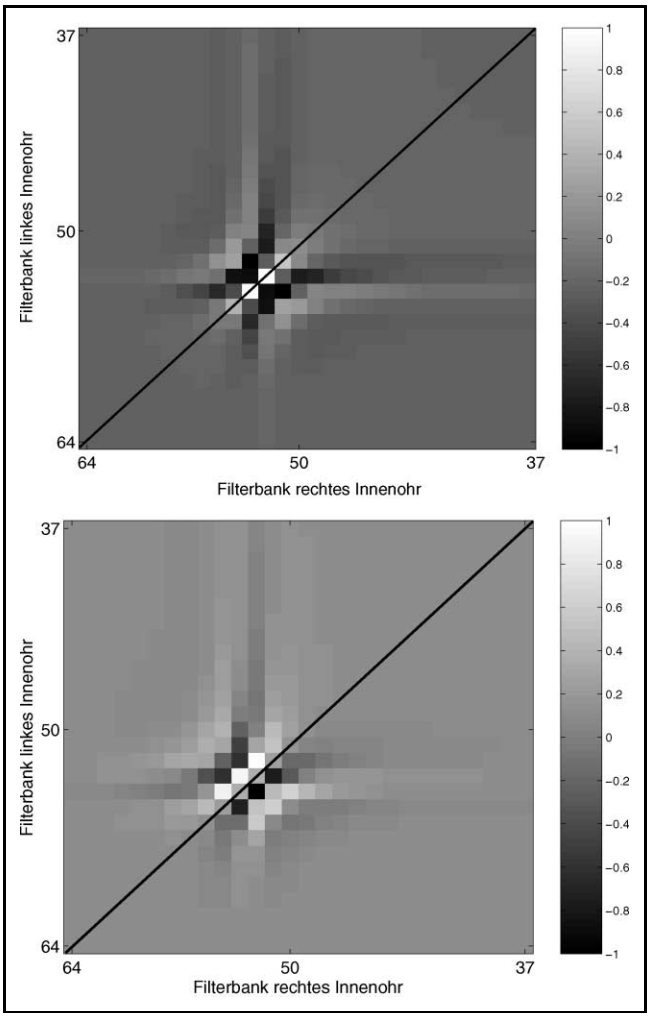

Bild 7: Stereagramm für eine Sinusschwingung von 300 Hz unter einem Einfallswinkel von $\alpha = 0°$ (oben) und $\alpha = 60°$ rechts (unten) im Bezug zum Roboterkopf.

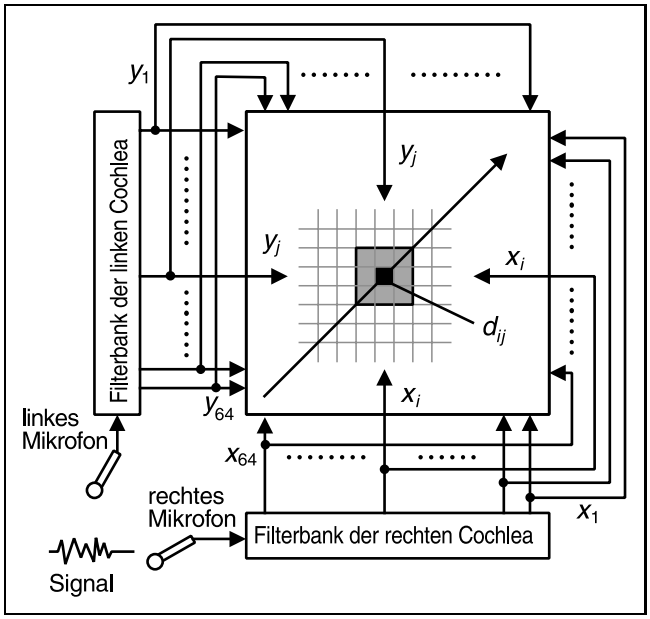

Bild 8: Comparausis Netzwerk zur Darstellung der Pegelunterschiede zwischen den einzelnen Kanälen der Filterbänke.

deutlicher Unterschied der Pegeldifferenzen festzustellen. Während bei einem Einfallswinkel $\alpha = 0°$ kein deutlicher Pegelunterschied zwischen den einzelnen Kanälen bemerkbar ist, sind klare Differenzen bei $\alpha = 90°$ zu erkennen, besonders auf der Hauptdiagonalen. Das Comparagramm beinhaltet somit wesentliche Merkmale der spezifischen Übertragungscharakteristik des Roboterkopfes. Das hier beschriebene System verwendet lediglich die Hauptdiagonale des Comparagramms zur Klassifikation der Einfallsrichtung des Signals anhand der Pegelunterschiede zwischen linkem und rechtem Ohr. Durch Hinzunahme weiterer Information aus dem Comparagramm ist eine Verbesserung der Klassifikation möglich, der Rechenaufwand ist allerdings entsprechend größer.

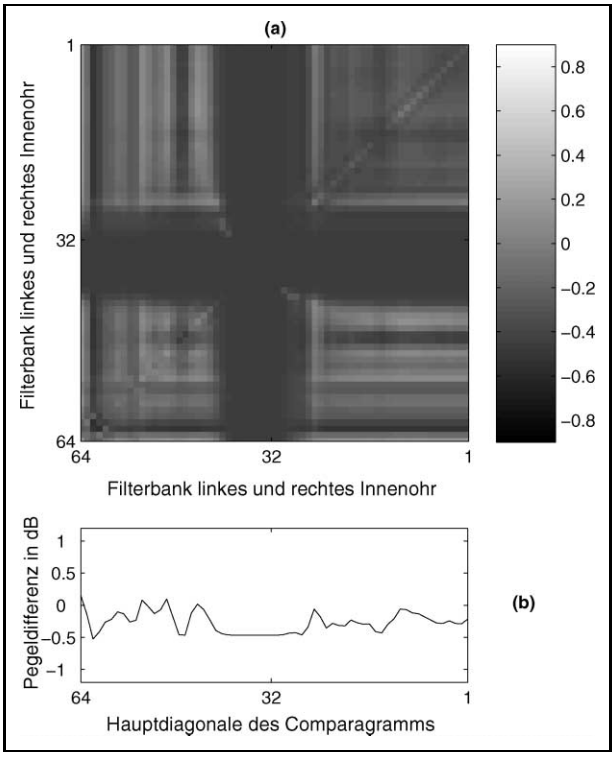

Bild 9: (**a**) Comparagramm bei einer Sinusschwingung von 1220 Hz unter einem Einfallswinkel von $\alpha = 0°$ im Bezug zum Roboterkopf. (**b**) Hauptdiagonale des Comparagramms.

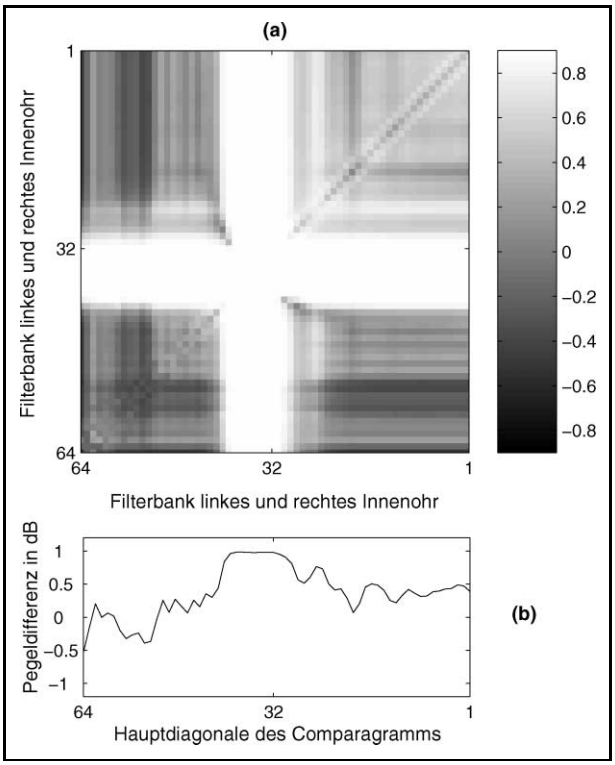

Bild 10: (**a**) Comparagramm bei einer Sinusschwingung von 1220 Hz unter einem Einfallswinkel von $\alpha = 90°$ links im Bezug zum Roboterkopf. (**b**) Hauptdiagonale des Comparagramms.

5 Merkmalsextraktion

Um den Einfallswinkel der Schallwelle in der horizontalen Ebene zu schätzen bzw. einer bestimmten Klasse zuzuordnen, muss man geeignete Merkmalsvektoren generieren, die die Informationen über interaurale Laufzeit- und Pegeldifferenzen eindeutig kodieren. Jede Klasse steht dabei für einen vorgegebenen Winkelsektor in der horizontalen Ebene. Die Merkmale für die Laufzeitunterschiede sind im Stereagramm und die Merkmale für die Pegeldifferenzen im Comparagramm enthalten. Die Extraktionsverfahren für die verschiedenen Merkmale, welche insgesamt den Merkmalsvektor bilden, werden im Folgenden erklärt.

Wie in den zwei Stereagrammen in Bild 7 zu sehen ist, verschiebt sich das Maximum (weiß) der Korrelation senkrecht zur Hauptdiagonalen auf die Nebendiagonalen mit zunehmendem Einfallswinkel α. Falls der Schall von links kommt, befindet sich das Maximum oberhalb der Hauptdiagonalen. Wenn er von rechts kommt, liegt das Maximum unterhalb der Hauptdiagonalen. Für den komplett betrachteten Winkelsektor von $-90°$ bis $90°$ (bezogen auf den frontalen Schalleinfall) verschiebt sich das Maximum aber nur im Bereich von der 1. oberen Nebendiagonalen bis zur 1. unteren Nebendiagonalen. Deswegen ist die Position des Maximums der Korrelation als Merkmal für die Zeitdifferenz zwischen beiden Signalen zu grob.

Um eine feinere Klassifizierung zu ermöglichen, müssen auch die anderen Nebendiagonalen zur Merkmalsextraktion herangezogen werden. Das Maximum der Korrelation entscheidet, welche Frequenz im Signal die größte Energie besitzt und ob das Signal von links oder rechts kommt. Die Elemente auf der Geraden, die senkrecht zur Hauptdiagonalen und durch das Korrelationsmaximum verlaufen, beinhalten eine genauere Information über die Zeitverschiebung. Bei einer Auflösung von 64 Kanälen pro Filterbank lohnt es sich nicht, mehr als drei Nebendiagonalen (nach oben und unten) zur Merkmalsgewinnung heranzuziehen, da kein weiterer Informationsgehalt im Bezug auf den Laufzeitunterschied in den weiteren Nebendiagonalen enthalten ist.

Der Ablauf der Merkmalsextraktion verläuft in drei Schritten. Zuerst sucht man das Korrelationsmaximum in der Haupt- bzw. der ersten oberen und unteren Nebendiagonale. Falls mehrere gleichwertige Maxima vorliegen,

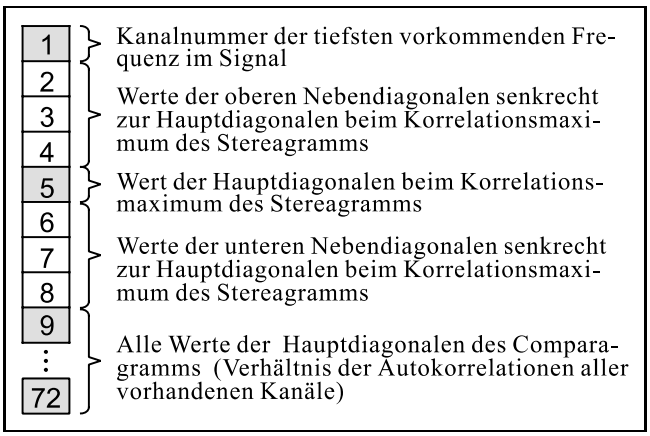

Bild 11: Zusammensetzung des Merkmalsvektors aus ITD- und ILD-Merkmalen.

wird das Maximum des Kanales mit der höchsten Kanalnummer (was der tiefsten Frequenz – bei harmonischen Signalen normalerweise der Grundschwingung – im Signal entspricht) gewählt. Dann können die Werte auf der Geraden senkrecht zur „Maximumsdiagonalen" als Merkmale definiert werden. Als letztes Merkmal benötigt man noch die Kanalnummer des Maximums, denn die Merkmale verändern sich nicht nur mit dem Einfallswinkel des Schalles, sondern auch mit seiner Frequenz.

Wie schon im vorherigen Kapitel beschrieben, spiegeln die Elemente der Hauptdiagonalen des Comparagramms die Pegeldifferenzen der beiden Kanäle gut wider. Deswegen werden alle Werte der Hauptdiagonalen des Comparagramms als Merkmale für die Pegelunterschiede benutzt. In Bild 11 ist die Zusammensetzung des gesamten Merkmalsvektors für ITD und ILD dargestellt.

6 Klassifikation mit einem PN-Netz

Für die Klassifikation der Schallquellen in der horizontalen Ebene wurde ein Probabilistisches Neuronales Netzwerk (PNN) mit zwei Schichten benutzt [16]. Die erste Schicht besteht aus Neuronen mit Bias, deren Aktivierungsfunktion die klassische Radial-Basis-Funktion, die Gauß-Funktion, ist. Die zweite Schicht des Netzes besteht aus einem Neuron mit einer Aktivierungsfunktion, die als Ausgang die Klasse mit der stärksten Aktivität der vorigen Schicht liefert. Damit bekommt man als Ergebnis immer die Klassifikationsklasse, zu der die Schallquelle – mit größter Wahrscheinlichkeit – gehört.

Das Netz bekommt normierte und unkorrelierte Eingangs- und Ausgangsdaten, die mehrmals zufällig in Lern- und Validierungsdaten im Verhältnis 3 zu 1 unterteilt werden. Die Leistung des gesamten Lokalisationsystems (Bild 12) ergibt sich dann aus der Mittelung der Ergebnisse verschiedener Lernvorgänge.

Die Eingangsdaten sind aufgenommene Signale, die aus reinen Sinusschwingungen, Überlagerungen von Harmonischen einer Grundfrequenz, Vokalen, einfachen Wörtern

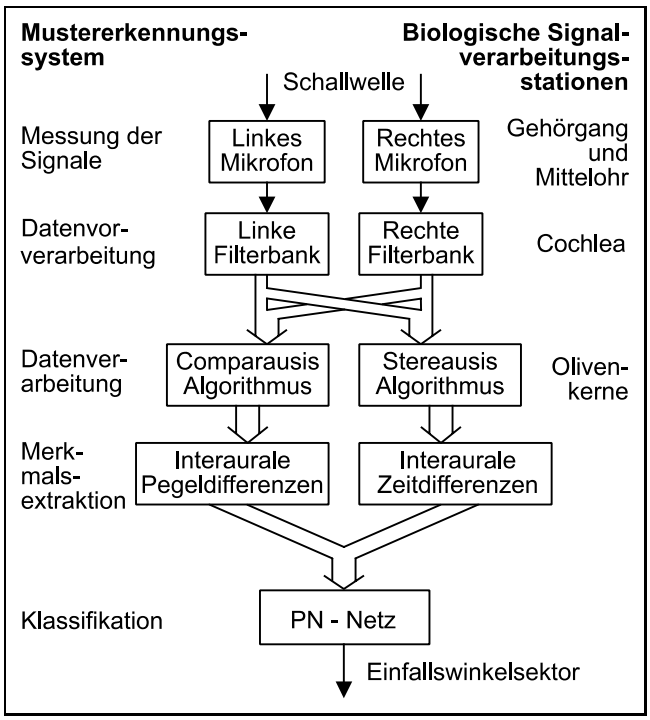

Bild 12: Lokalisationssystem, biologisches Vorbild und technische Umsetzung für Schallaufnahmen in echoarmer Umgebung.

und Klatschgeräuschen bestehen. Die Signale wurden aus verschiedenen Richtungen, die in sieben Winkelsektoren unterteilt sind, aufgenommen. Sieben verschiedene Klassen wurden definiert, von −90° bis zu 90° in 30° Schritten. Der Einfallswinkel jeder Schallquelle wird einer dieser Klassen zugeordnet und damit lokalisiert. Dabei wurde die Frequenz der Sinusschwingungen schrittweise von 40 Hz bis zu 1,2 kHz erhöht. Aufgenommen wurden die Signale mit dem ganzen System, einschließlich Roboterkopf und Mikrofonen, in einem echoarmen Raum. Dabei hatten die Mikrofone einen Abstand von 28 cm und standen in einem Winkel von 120° zueinander.

7 Ergebnisse

Benutzt man nur die ITD-Merkmale aus dem Stereausis Algorithmus, dann ist die Leistung des Systems bei den mittleren Sektoren (−30° bis zu 30°) befriedigend und liegt im Durchschnitt bei etwa 57% (Bild 13). Andererseits werden seitlich einfallende Signale mit sehr kleiner Wahrscheinlichkeit, etwa 27%, korrekt zugeordnet.

Benutzt man auch die ILD-Merkmale aus dem Comparausis Algorithmus, so wird die Leistung des Systems verbessert, besonders bei seitlich einfallenden Schallquellen (Bild 14). Die durchschnittliche Wahrscheinlichkeit einer korrekten Zuordnung liegt jetzt bei 63% in allen Sektoren. Auch die Verwechslung beschränkt sich nun auf benachbarte Winkelsektoren.

Dabei leisten sowohl die ITD- als auch die ILD-Merkmale einen Anteil zur Klassifikation. Die ITD-Merkmale können bei kleinen Einfallswinkeln eindeutige Unterschiede

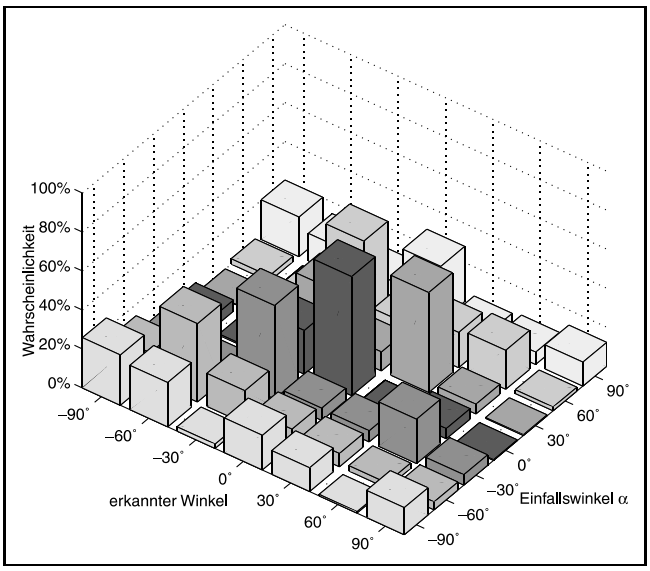

Bild 13: Konfusionsmatrix für Schallaufnahmen in echoarmer Umgebung (nur ITD-Merkmale). Die Hauptdiagonale beinhaltet die fehlerfreien Erkennungen.

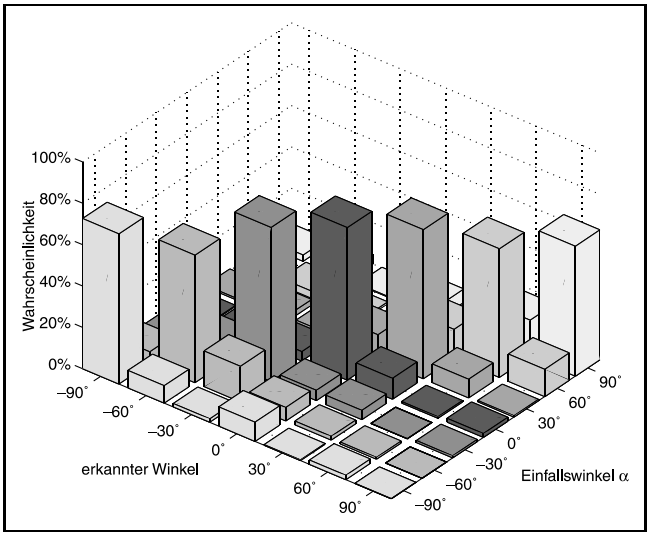

Bild 14: Konfusionsmatrix für Schallaufnahmen in echoarmer Umgebung (ITD- und ILD-Merkmale).

besser kodieren. Andererseits trennen die ILD-Merkmale bei großen Einfallswinkeln die Winkelsektoren genauer, weil die Intensitätsunterschiede durch die Dämpfung des Roboterkopfes größer sind und damit entsprechend bessere ILD-Merkmale vom Comparausis-Algorithmus geliefert werden können. Zusammen bieten die ITD- und ILD-Merkmale eine gute Grundlage für die Lokalisation von Schallquellen über das gesamte Sichtfeld des Roboters. Eine wiederholte Aktivität der aus einer Richtung sendenden Schallquelle, z. B. ein menschlicher Sprecher, verbessert die Erkennungswahrscheinlichkeit zu $P = 1 - 0{,}37^n$, wobei n die Anzahl der gesendeten Signale ist.

Mit zunehmender Anzahl der Filter jeder Filterbank und damit auch der Merkmale, ist eine weitere Verbesserung der Klassifikationsgüte möglich, allerdings muss dabei auch die Zunahme der Rechenzeit berücksichtigt werden.

8 Zusammenfassung und Ausblick

Das in dieser Arbeit entworfene, biologisch inspirierte System zur Lokalisation von Schallquellen ist eines der ersten Systeme für das Richtungshören von Robotern und bietet eine solide Grundlage für weitere Untersuchungen und Verbesserungen auf dem Gebiet des binauralen Richtungshörens von Robotern. Die Lokalisationsgüte der Schallquellen im Sichtfeld des Roboterkopfes ist zwar gut, sie kann aber mit einigen Erweiterungen des Systems noch verbessert werden. Insbesondere ist dies für den Fall von verhallten Räumen erforderlich.

Literatur

[1] B. Kollmeier: *Die Oldenburger Hörforschung, Einblicke, Carl von Ossietzky Universität Oldenburg, Nr. 33, S. 4–7, 2001*.

[2] G. Ehret, R. Romand: *The central auditory system, Oxford University Press, 1997*.

[3] R.O. Duda, C. Lim: *Estimating the azimuth and elevation of a sound source from the output of a cochlear model, IEEE Proc. ASILOMAR-28, Conf. on Signals, Systems and Computers, pp. 399–403, 1994*.

[4] R.O. Duda, W. Chau: *Combined monaural and binaural localization of sound sources, IEEE Proc. ASILOMAR-29, pp. 1281–1285, 1996*.

[5] K. Martin: *Estimating azimuth and elevation from interaural differences, Proc. IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 96–99, 1995*.

[6] J. Blauert: *Spatial hearing: The psychophysics of human sound localization, MIT Press, 1996*.

[7] S.A. Shamma, N. Shen, P. Gopalaswamy: *Stereausis: Binaural processing without neural delays, Journal of Acoustic Society Am. 86, pp. 989–1006, 1989*.

[8] S.A. Shamma: *On the Role of space and time in auditory processing, TRENDS in Cognitive Science, Vol. 5 No. 8, pp. 340–348, 2001*.

[9] E.R. Kandel, J.H. Schwartz, T.M. Jessel: *Neurowissenschaften, eine Einführung, Spektrum Akademischer Verlag, 1996*.

[10] G. Langner: *Das Phantomgeräusch Tinnitus ist eine Störung der Informationsverarbeitung im Gehirn, thema Forschung, Nr. 1, S. 124–131, 1998*.

[11] L.O. Douglas: *Ascending efferent projections of the Superior Olivary Complex, Microscopy Research and Technique, Vol.51, pp. 340–348, 2001*.

[12] G. Langner, M. Sams, P. Heil, H. Schulze: *Frequency and periodicity are represented in orthogonal maps in the human auditory cortex: evidence from magnetoencephalography, Journal Comp. Physiol., Vol. 181, pp. 665–676, 1997*.

[13] L.A. Jefress: *A place theory of sound localization, Journal of Comparative and Physiological Psychology, Vol. 41, pp. 35–39, 1948*.

[14] R. Patterson, I. Nimmo-Smith, J. Holdsworth, P. Rice: *Spiral VOS final report: Part A, the auditory filterbank, Internal Report, University of Cambridge, 1988*.

[15] M. Slaney: *An efficient Implementation of the Patterson-Holdsworth auditory filter bank, Apple Computer Technical Report 35, 1993*.

[16] P.D. Wasserman: *Advanced methods in neural computing, Van Nostrand Reinhold, pp. 35–55, 1993.*

Manuskripteingang: 8. Dezember 2002.

Prof. Dr.-Ing. Jürgen Adamy ist Leiter des Fachgebietes Regelungstheorie und Robotik der Technischen Universität in Darmstadt. Hauptarbeitsgebiete: Regelungsverfahren, Computational Intelligence, Bionik und autonome mobile Roboter.

Adresse: Technische Universität Darmstadt, Fachgebiet Regelungstheorie und Robotik, Landgraf-Georg-Str. 4, D-64283 Darmstadt, Fax: 0049/6151/16-2507, E-Mail: Jadamy@iat.tu-darmstadt.de

Dipl.-Ing. Kyriakos Voutsas ist Wissenschaftlicher Mitarbeiter des Fachgebietes Regelungstheorie und Robotik der Technischen Universität in Darmstadt. Hauptarbeitsgebiete: Bio-inspired systems, Sensorik für Roboter, Artificial Neural Networks.

Adresse: Technische Universität Darmstadt, Fachgebiet Regelungstheorie und Robotik, Petersenstr. 20, D-64287 Darmstadt, Fax: 0049/6151/16-7424, E-Mail: kvoutsas@rtr.tu-darmstadt.de

Dipl.-Ing. Volker Willert ist Wissenschaftlicher Mitarbeiter des Fachgebietes Regelungstheorie und Robotik der Technischen Universität in Darmstadt. Hauptarbeitsgebiete: Bildverarbeitung, Bio-inspired systems.

Adresse: Technische Universität Darmstadt, Fachgebiet Regelungstheorie und Robotik, Landgraf-Georg-Str. 4, D-64283 Darmstadt, Fax: 0049/6151/16-2507, E-Mail: volker@rtr.tu-darmstadt.de